The MetaArchive of Southern Digital Culture

Building a Multi-Institutional Digital Preservation Network

Tyler Walters
Georgia Institute of Technology Library

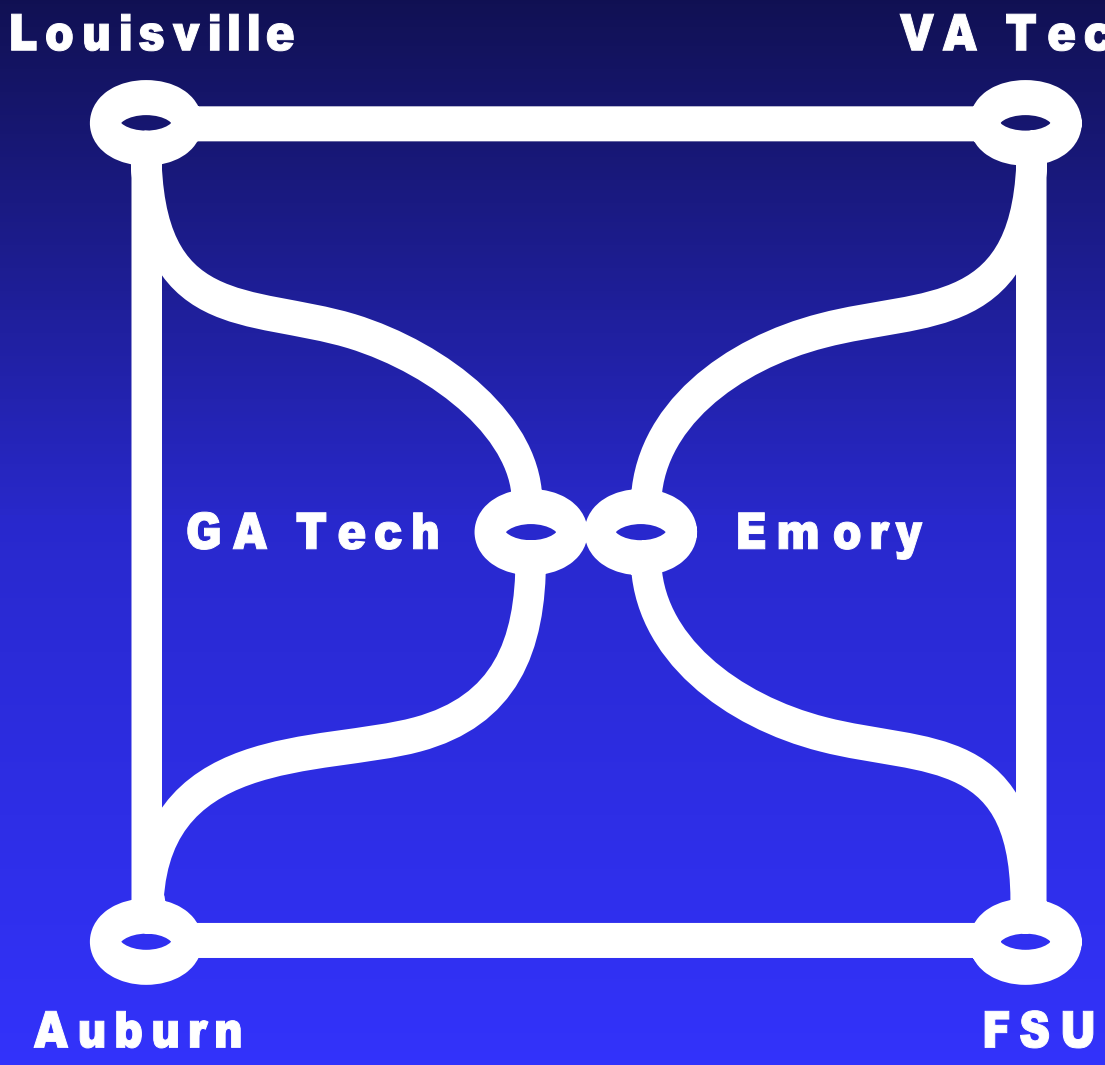
National Conference on Digital Government Research
Atlanta, GA, May 16, 2005

MetaArchive Project Summary

- Six partner institutions will collaborate with LoC on a three year $1.4M effort to develop a cooperative for the preservation of digital content
- Content focus:

  Southern culture and history

MetaArchive
Preservation Network Partners



Project Goals

1. Create a conspectus of digital content within the subject domain held by the partner sites
2. Harvested body of the most critical content to be preserved (3 terabytes, w/ capability to expand)
3. Develop a model cooperative agreement for ongoing collaboration and sustainability
4. Distributed preservation network infrastructure based on the LOCKSS software

Key Preservation Network Features

1. Distributed Preservation Strategy
2. Flexible Organizational Model
3. Formal Content Selection Process
4. Capability for Migrating Archives
5. Dark Archiving Strategy
6. Low Cost to Deployment
7. Self-Sustaining Incentives
8. Simple Preservation Exchange Mechanisms with the Library of Congress

MetaArchive Preservation Network

- Preservation Network Approach:
- **Develop distributed preservation network infrastructure for shared preservation of digital content (many locations, many copies over time)**
- **Peer-to-peer network architecture for content preservation. Six nodes. 3 TB storage in aggregate. Each node communicates with all other nodes**
- **All nodes serve as joint custodians of content harvested. No data lost if one node withdraws or becomes dysfunctional. Reliably preserved and validated at all preservation sites**

Adapting LOCKSS Software

- Allows Cooperative to practice its “distributed digital replication” approach. Replicate archival material, not just published material, as originally designed:
  - data integrity checks
  - rigorous security checks
  - focused web crawls to gather/ingest digital content
  - problem of dynamically constructed content (to be studied)

Dark Archiving & Low Costs

- Advantage: many preservation efforts mix high accessibility online with long-term access (preservation). High accessibility = high costs
- Network's preserved content discoverable via metadata (OAI-PMH), downloadable restricted means
- Will maintain publicly accessible registry of collections / items preserved
- Designed for minimal expenditures, low barriers to adoption, for medium-sized institutions
- Runs on inexpensive computers, modest degree of systems administration for ongoing maintenance

Hardware (all universities)

- SAN array:
  - Dell/EMC AX100 Array (single processor)
    - 3TB storage space (Four 3X250 GB 7200 rpm serial ATA hard disk drives)
- SAN server:
  - Dell PowerEdge 1850 (2 processors)
    - 3.0GHz/1MB Cache, Xeon 800MHz
- Firewall
  - Dell PowerEdge 1850 (1 processor)
    - 3.0Ghz/1MB Cache, Xeon 800MHz
  - Dell PowerConnect 2616 Unmanaged Switch

Content Selection

- Preservation efforts are likely to be most coherent around shared focus
- Selection of collections to be preserved made by teams of subject specialist librarians and archivists at contributing institutions
- These teams are creating a conspectus of collections for consideration and prioritization
- Using collection framework of the *Encyclopedia of Southern Culture*

Conspectus Metadata Schema

- Informed by many current, ongoing efforts:
  - Dublin Core Collection Description Application Profile
  - UIUC IMLS Collection Description Metadata Schema
  - UKOLN RSLP Collection Description Schema
  - Western States Dublin Core Metadata Best Practice
- MetaArchive conspectus schema has been developed and is available on the project website:

(http://www.MetaArchive.org)

Cooperative Agreement

- **Will develop a simple and flexible cooperative agreement as a model for other institutions seeking to cooperate in digital preservation:**
  - **Membership criteria (and withdrawal)**
  - **Roles and responsibilities – joint and equal custodians of the content harvested**
  - **Sustainability plan (over time)**
  - **Ensure broad applicability**
  - **Self-sustaining incentive -- preserve institution-produced content**

* **All partners are members of Internet 2 Consortium**

Types of Content

- **Website exhibitions**
- **Digital masters (of scanned images of brittle analog originals where "rescanning" opportunities are low)**
- **"Research databases" (created by scholars working in the field, i.e. ethnographic investigations, image databases, digitally recorded interviews, etc.).**
- **Institutional materials – digital collections, digital intellectual output, i.e. reports, papers, learning objects, etc.**

- Questions / Comments:

Tyler Walters
Associate Director, Technology and Resource Services
Georgia Tech Library & Information Center

404-385-4489
Tyler.Walters@library.gatech.edu

http://www.MetaArchive.org